Introduction to Machine Learning

for the Life Sciences

JJ Valletta

February 24, 2015

Housekeeping

- Who are we?
- Timetable
- Important Info
- Contact emails

www.exeter.ac.uk/as/rdp/

Workshop learning outcomes

- Understand the key concepts and terminology used in the field of machine learning
- Apply machine learning algorithms in R and apply them to your own datasets
- Recognise practical issues in data analysis

Overview

- What is machine learning?
- Types of learning methods
- Statistics vs Machine Learning
- Formulating a machine learning problem
- Terminology
- Applications in life sciences

Overview

- What is machine learning?
- Types of learning methods
- Statistics vs Machine Learning
- Formulating a machine learning problem
- Terminology
- Applications in life sciences

Overview

- What is machine learning?
- Types of learning methods
- Statistics vs Machine Learning
- Formulating a machine learning problem
- Terminology
- Applications in life sciences

Overview

- What is machine learning?
- Types of learning methods
- Statistics vs Machine Learning
- Formulating a machine learning problem
- Terminology
- Applications in life sciences

Overview

- What is machine learning?
- Types of learning methods
- Statistics vs Machine Learning
- Formulating a machine learning problem
- Terminology
- Applications in life sciences

Overview

- What is machine learning?
- Types of learning methods
- Statistics vs Machine Learning
- Formulating a machine learning problem
- Terminology
- Applications in life sciences

Overview

- What is machine learning?
- Types of learning methods
- Statistics vs Machine Learning
- Formulating a machine learning problem
- Terminology
- Applications in life sciences

What my mum thinks machine learning is

Who uses machine learning?

Who uses machine learning?

Machine Learning in Ecosystem Informatics and Sustainability

Thomas G. Dietterich
School of Electrical Engineering and Computer Science
Oregon State University
tgd@cs.orst.edu

MACHINE LEARNING IN THE LIFE SCIENCES

©BRAND X, PHOTODISC

Machine Learning in the Life Sciences

How it is Used on a Wide Variety of Medical Problems and Data

KRZYSZTOF J. CIOS, LUKASZ A. KURGAN, AND MAREK REFORMAT

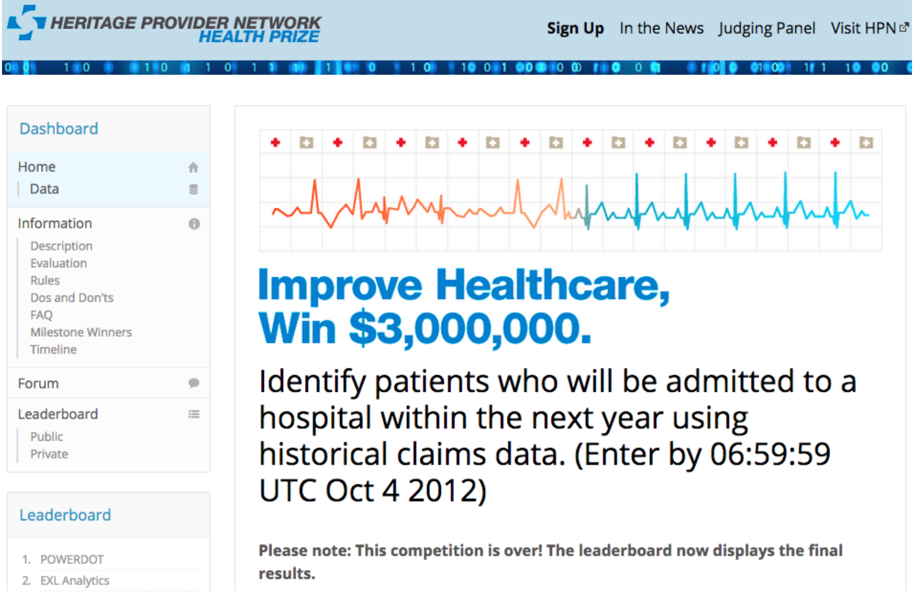
VOLUME 83, NO. 2 *THE QUARTERLY REVIEW OF BIOLOGY* JUNE 2008

MACHINE LEARNING METHODS WITHOUT TEARS: A PRIMER FOR ECOLOGISTS

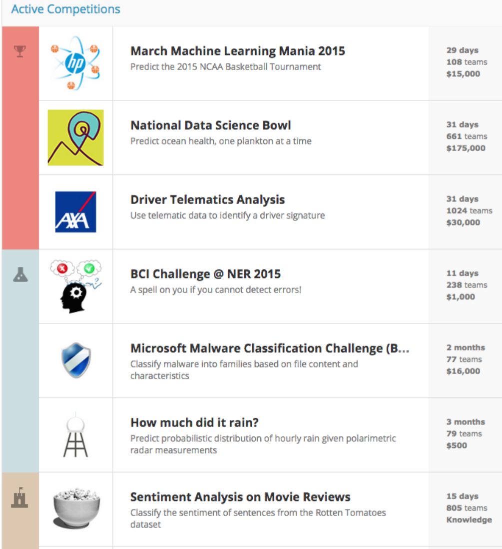
Data Analysis and Mining in the Life Sciences

Nam Huyn
SurroMed, Inc.
2375 Garcia Ave, Mountain View, CA 94043, USA
phuyn@surromed.com

There are even competitions now!



Lots of them actually...



Source: http://www.kaggle.com/

So what is machine learning?

So what is machine learning?

> A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E
>
> — Tom Mitchell

So what is machine learning?

A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E

— Tom Mitchell

A scientific discipline that explores the construction and study of algorithms that can learn from data. Such algorithms operate by building a model from example inputs and using that to make predictions or decisions, rather than following strictly static program instructions.

— Wikipedia

So what is machine learning?

A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E

— Tom Mitchell

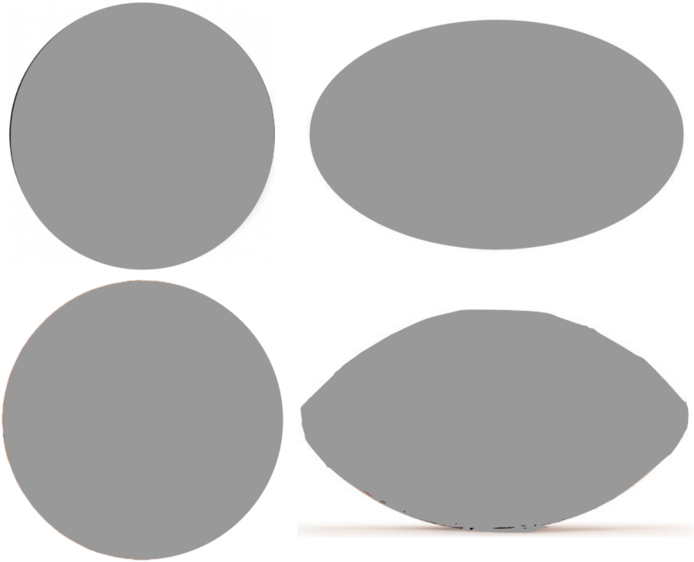
A scientific discipline that explores the construction and study of algorithms that can learn from data. Such algorithms operate by building a model from example inputs and using that to make predictions or decisions, rather than following strictly static program instructions.

— Wikipedia

Machines learn using flashcards

— JJ

Group by shape (unsupervised learning)



Add labels (supervised learning)

Types of learning methods

Unsupervised learning: Inputs have *no* corresponding output labels

- **Clustering** - discovering groups having similar attributes (e.g unlabelled gene expression profiles)
- **Density Estimation** - determine probability distribution of data (e.g species distribution model)
- **Dimensionality Reduction** - identify and remove redundant dimensions

Supervised learning: Inputs have corresponding output labels

- **Regression** - output is a continuous variable (e.g CO_2 *concentration ppm*)
- **Classification** - output is categorical (e.g benign (0) or cancerous (1) tumour)

Reinforcement learning: Finding suitable actions to maximise a reward by a process of trial and error. Tradeoff between exploration (use new actions) vs exploitation (use actions already known to work)

Types of learning methods

Unsupervised learning: Inputs have *no* corresponding output labels

- **Clustering** - discovering groups having similar attributes (e.g unlabelled gene expression profiles)
- **Density Estimation** - determine probability distribution of data (e.g species distribution model)
- **Dimensionality Reduction** - identify and remove redundant dimensions

Supervised learning: Inputs have corresponding output labels

- **Regression** - output is a continuous variable (e.g CO_2 *concentration ppm*)
- **Classification** - output is categorical (e.g benign (0) or cancerous (1) tumour)

Reinforcement learning: Finding suitable actions to maximise a reward by a process of trial and error. Tradeoff between exploration (use new actions) vs exploitation (use actions already known to work)

Types of learning methods

Unsupervised learning: Inputs have *no* corresponding output labels

- **Clustering** - discovering groups having similar attributes (e.g unlabelled gene expression profiles)
- **Density Estimation** - determine probability distribution of data (e.g species distribution model)
- **Dimensionality Reduction** - identify and remove redundant dimensions

Supervised learning: Inputs have corresponding output labels

- **Regression** - output is a continuous variable (e.g CO_2 *concentration ppm*)
- **Classification** - output is categorical (e.g benign (0) or cancerous (1) tumour)

Reinforcement learning: Finding suitable actions to maximise a reward by a process of trial and error. Tradeoff between exploration (use new actions) vs exploitation (use actions already known to work)

Types of learning methods

Unsupervised learning: Inputs have *no* corresponding output labels

- **Clustering** - discovering groups having similar attributes (e.g unlabelled gene expression profiles)
- **Density Estimation** - determine probability distribution of data (e.g species distribution model)
- **Dimensionality Reduction** - identify and remove redundant dimensions

Supervised learning: Inputs have corresponding output labels

- **Regression** - output is a continuous variable (e.g CO_2 *concentration ppm*)
- **Classification** - output is categorical (e.g benign (0) or cancerous (1) tumour)

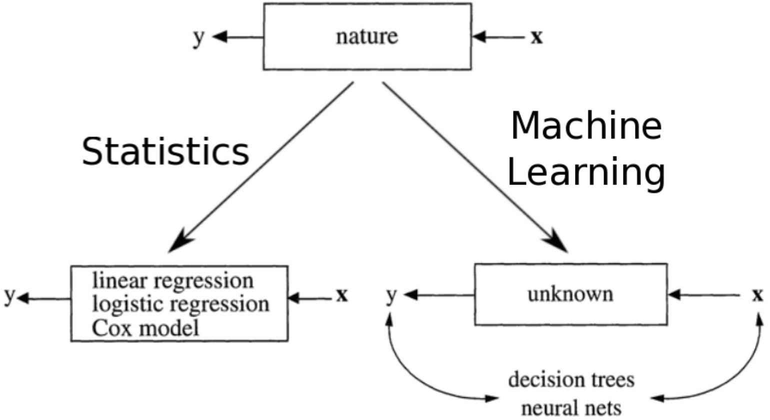
Reinforcement learning: Finding suitable actions to maximise a reward by a process of trial and error. Tradeoff between exploration (use new actions) vs exploitation (use actions already known to work)

Statistics vs Machine Learning

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman



Statistics vs Machine Learning

Statistics

- **Objective** - provide humans a set of data analysis tools
- **Focus** - a probabilistic data model and its interpretability
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Objective** - replace humans in the processing of data
- **Focus** - algorithm that achieves excellent prediction accuracy
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning

Statistics

- **Objective** - provide humans a set of data analysis tools
- **Focus** - a probabilistic data model and its interpretability
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Objective** - replace humans in the processing of data
- **Focus** - algorithm that achieves excellent prediction accuracy
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning

Statistics

- **Objective** - provide humans a set of data analysis tools
- **Focus** - a probabilistic data model and its interpretability
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Objective** - replace humans in the processing of data
- **Focus** - algorithm that achieves excellent prediction accuracy
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning

Statistics

- **Objective** - provide humans a set of data analysis tools
- **Focus** - a probabilistic data model and its interpretability
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Objective** - replace humans in the processing of data
- **Focus** - algorithm that achieves excellent prediction accuracy
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning

Statistics

- **Objective** - provide humans a set of data analysis tools
- **Focus** - a probabilistic data model and its interpretability
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Objective** - replace humans in the processing of data
- **Focus** - algorithm that achieves excellent prediction accuracy
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning

Statistics

- **Objective** - provide humans a set of data analysis tools
- **Focus** - a probabilistic data model and its interpretability
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Objective** - replace humans in the processing of data
- **Focus** - algorithm that achieves excellent prediction accuracy
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning

Statistics

- **Objective** - provide humans a set of data analysis tools
- **Focus** - a probabilistic data model and its interpretability
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Objective** - replace humans in the processing of data
- **Focus** - algorithm that achieves excellent prediction accuracy
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning

This enterprise (statistics) has at its heart the belief that a statistician, by imagination and by looking at the data, can invent a reasonably good parametric class of models for a complex mechanism devised by nature. Then parameters are estimated and conclusions are drawn. The conclusions are about the model's mechanism, and not about nature's mechanism!

— Leo Breiman

The best solution could be an algorithmic model (machine learning), or maybe a data model, or maybe a combination. But the trick to being a scientist is to be open to using a wide variety of tools.

— Leo Breiman

Formulating a machine learning problem

1. Question/Hypothesis (start generic then narrow it down)
2. Gather useful data
3. **Extract features** (*most important step*)
4. Choose a machine learning algorithm (*probably least critical step*)
5. Build a predictive model
6. Validate model
7. Select model having the best generalisation capabilities

Formulating a machine learning problem

1. Question/Hypothesis (start generic then narrow it down)
2. Gather useful data
3. **Extract features** (*most important step*)
4. Choose a machine learning algorithm (*probably least critical step*)
5. Build a predictive model
6. Validate model
7. Select model having the best generalisation capabilities

Formulating a machine learning problem

1. Question/Hypothesis (start generic then narrow it down)
2. Gather useful data
3. **Extract features** (*most important step*)
4. Choose a machine learning algorithm (*probably least critical step*)
5. Build a predictive model
6. Validate model
7. Select model having the best generalisation capabilities

Formulating a machine learning problem

1. Question/Hypothesis (start generic then narrow it down)
2. Gather useful data
3. **Extract features** (*most important step*)
4. Choose a machine learning algorithm (*probably least critical step*)
5. Build a predictive model
6. Validate model
7. Select model having the best generalisation capabilities

Formulating a machine learning problem

1. Question/Hypothesis (start generic then narrow it down)
2. Gather useful data
3. **Extract features** (*most important step*)
4. Choose a machine learning algorithm (*probably least critical step*)
5. Build a predictive model
6. Validate model
7. Select model having the best generalisation capabilities

Formulating a machine learning problem

1. Question/Hypothesis (start generic then narrow it down)
2. Gather useful data
3. **Extract features** (*most important step*)
4. Choose a machine learning algorithm (*probably least critical step*)
5. Build a predictive model
6. Validate model
7. Select model having the best generalisation capabilities

Formulating a machine learning problem

1. Question/Hypothesis (start generic then narrow it down)
2. Gather useful data
3. **Extract features** (*most important step*)
4. Choose a machine learning algorithm (*probably least critical step*)
5. Build a predictive model
6. Validate model
7. Select model having the best generalisation capabilities

Formulating a machine learning problem

1. Question/Hypothesis (start generic then narrow it down)
2. Gather useful data
3. **Extract features** (*most important step*)
4. Choose a machine learning algorithm (*probably least critical step*)
5. Build a predictive model
6. Validate model
7. Select model having the best generalisation capabilities

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonous model i.e a model which is complex enough to describe "well" our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generlisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonous model i.e a model which is complex enough to describe "well" our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generlisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonous model i.e a model which is complex enough to describe "well" our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generlisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonous model i.e a model which is complex enough to describe "well" our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generlisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonous model i.e a model which is complex enough to describe “well” our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generlisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonous model i.e a model which is complex enough to describe “well” our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generlisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonous model i.e a model which is complex enough to describe “well” our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generlisation error

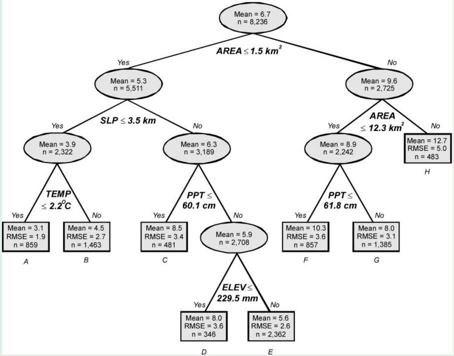
Applications in life sciences

Predicting fish species richness

Olden *et al.* Q Rev Biol 2008, 83(2):171-193

Features: Lake surface area, shoreline perimeter, air temperature, precipitation and elevation

Method: Decision trees (supervised)



Applications in life sciences

Detection of malarial parasites

Purwar *et al.* Malar J 2011, 10:364

Features: Image intensity

Method: Modified k-means clustering (unsupervised)

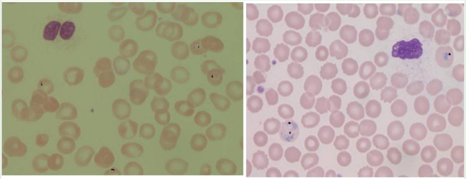
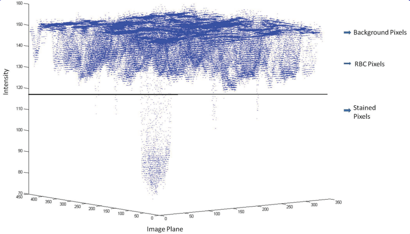


Figure 17 Parasites marked image.



Applications in life sciences

Creating carbon-density maps

Baccini *et al.* Nature Clim. Change 2012, 2:182-185

Features: Light detection and ranging (LiDAR) (elevation data)

Method: Random forests (ensemble of decision trees) (supervised)

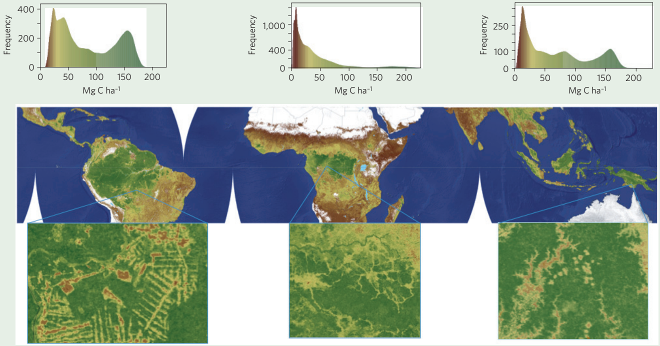


Figure 1 | Carbon contained in the aboveground live woody vegetation of tropical America, Africa and Asia (Australia excluded). The upper panels show the frequency distribution of carbon in units of Mg C ha^{-1} for each region. Inset figures across the bottom provide higher-resolution examples of the spatial detail present in the satellite-derived biomass data set. Carbon amount is represented in the maps as a colour scheme from dark brown (low carbon) to dark green (high carbon). See upper panels for numeric values.

Applications in life sciences

Acoustic classification of multiple simultaneous bird species

Briggs *et al.* J Acoust Soc Am 2012, 131(6):4640-4650

Features: Segments in spectrogram (time vs frequency) from 10 secs audio recordings (corresponding to syllables of bird call)

Method: Multi-instance multi-label learning (supervised)

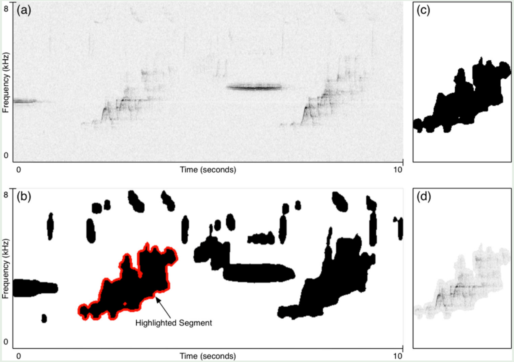


FIG. 3. (Color online) Extracting a syllable from the segmentation results. (a) The original spectrogram, (b) the binary mask generated by our segmentation algorithm. The highlighted segment will be further processed in this example. Note that several other segments overlap in time. (c) A cropped mask of the highlighted segment. (d) The masked and cropped spectrogram corresponding to the highlighted segment.